

Bandwith Selection for Kernel Distribution Function Estimation

Naomi Altman
Biometrics Unit
Cornell University
Ithaca, New York

Christian Léger
Département de mathématiques
et de statistique
Université de Montréal
Montréal, Québec

BU-1246-M

July 1994

Bandwidth Selection for Kernel Distribution Function Estimation

| | |
|--------------------|------------------------------|
| Naomi Altman * | Christian Léger † |
| Biometrics Unit | Département de mathématiques |
| Cornell University | et de statistique |
| | Université de Montréal |

July 1994

Abstract

Leave-one-out cross-validation is a popular and readily implemented heuristic for bandwidth selection in nonparametric smoothing problems. In this note we elucidate the role of leave-one-out selection criteria by discussing a criterion introduced by Sarda (1993) for bandwidth selection for kernel distribution function estimators (KDFEs). We show that for this problem, use of the leave-one-out KDE in the selection procedure is asymptotically equivalent to leaving none out. This contrasts with kernel density estimation, where use of the leave-one-out density estimator in the selection procedure is critical.

Unfortunately, simulations show that neither method works in practice, even for samples of size as large as 1000. In fact, we show that for any fixed bandwidth, the expected value of the derivative of the leave-one-out criterion is asymptotically positive. This result and our simulations suggest that the criteria are increasing and that for sufficiently large samples (e.g., $n = 100$), the smallest available bandwidth will always be selected, thus contradicting the optimality result of Sarda for this estimator.

As an alternative to minimizing a selection criterion, we propose a plug-in estimator of the asymptotically optimal bandwidth. Simulations suggest that the plug-in is a good esti-

*Supported by Hatch Grant 151410 NYF

†Supported by NSERC (Canada) and FCAR (Québec)

mator of the asymptotically optimal bandwidth even for samples as small as 10 observations and is not too far from the finite sample optimal bandwidth.

AMS subject classification: Primary 62 G05 Secondary 62 G20

Keywords and phrases: Distribution function; nonparametric estimation; smoothing parameter selection; cross-validation; leave-one-out estimator.

1 Introduction

Leave-one-out cross-validation is a popular and readily implemented heuristic for bandwidth selection in nonparametric smoothing problems. In this note we elucidate the role of leave-one-out selection criteria by discussing a criterion introduced by Sarda (1993) for bandwidth selection for kernel distribution function estimators (KDFEs). We show that for this problem, use of the leave-one-out KDE in the selection procedure is asymptotically equivalent to leaving none out.

Unfortunately, simulations show that neither method works in practice, even for samples of size as large as 1000. In fact, we show that for any fixed bandwidth, the expected value of the derivative of the leave-none-out criterion is asymptotically positive. This result and our simulations suggest that the criteria are increasing and that for sufficiently large samples (e.g., $n = 100$), the smallest available bandwidth will always be selected, thus contradicting Sarda's optimality result for this selector.

Sarda's selection criterion evaluates the KDE as a function of the bandwidth, by evaluating against another estimator of the distribution function, the empirical distribution function (EDF). Similarly, Müller, Stadtmüller, and Schmitt (1987) suggested selecting the bandwidth of a kernel estimator of the derivative of a nonparametric regression function by evaluating against a divided difference estimator of the derivative which does not depend on the bandwidth. Leave-some-out estimation arises in this context because the covariance between the kernel estimator and the proxy for the target function introduces a term depending on the

bandwidth which is of the same order of magnitude as the risk of the kernel estimator. In the context of regression derivatives the covariance can be eliminated by using disjoint sets to evaluate the kernel derivative estimator and the divided difference which is the proxy for the derivative function and the same proves true in the current problem. The EDF used in Sarda's criterion uses all the data. Leaving some out of the KDE does not provide the appropriate correction to the selection criterion, because the remaining terms are still correlated with the EDF. Using disjoint subsets of the data to estimate the KDE and EDF does eliminate the covariance term. However, adequate estimation of the proxy function, the EDF, requires a substantial fraction of the data, so that the selected bandwidth needs a sample size adjustment. The computational burden is also high.

As an alternative to the use of a bandwidth selection criterion for KDE, we introduce a plug-in estimator of the asymptotically optimal bandwidth. This estimator is similar to plug-in estimators for bandwidth selection in density estimation (Jones, Marron and Park, 1991; Park and Marron, 1991; Sheather and Jones, 1991). We show that the estimator is consistent and performs well in simulations.

We describe the KDE and the estimators of average squared error in the next section. We show the asymptotic equivalence of the leave-one-out and leave-none-out criteria and show that the expected value of the derivative of the leave-none-out criterion is asymptotically positive. We also show how the problems of bandwidth selection for density and distribution function estimation differ. In Section 3, we introduce the plug-in estimator of the optimal bandwidth. Section 4 contains simulation results that show that the methods which minimize leave- n -out estimators of risk, $n = 0, 1$, are not useful in practice, while the plug-in method does well. An appendix contains the proofs.

2 The leave-none-out procedure

Let X_1, \dots, X_n be identically and independently distributed from distribution function F . The kernel distribution function estimator \hat{F}_h was introduced by Nadaraya (1964) and is defined by

$$\hat{F}_h(x) = n^{-1} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right),$$

where K is the distribution function of a positive kernel k , (i.e. $K(x) = \int_{-\infty}^x k(t) dt$) and h is the bandwidth.

Two measures of quality of the kernel estimator are the Integrated Squared Error and the Mean Integrated Squared Error

$$\begin{aligned} ISE(h) &= \int (\hat{F}_h(x) - F(x))^2 W(x) dF(x) \\ MISE(h) &= E \int (\hat{F}_h(x) - F(x))^2 W(x) dF(x), \end{aligned}$$

where W is a nonnegative weight function. A discrete approximation to $MISE$ is the Average Squared Error

$$ASE(h) = n^{-1} \sum_{i=1}^n [\hat{F}_h(X_i) - F(X_i)]^2 W(X_i).$$

This random variable cannot be computed since F is unknown. But F can be replaced by the EDF F_n leading to what we call a leave-none-out estimator of $ASE(h)$,

$$LNO(h) = n^{-1} \sum_{i=1}^n [\hat{F}_h(X_i) - F_n(X_i)]^2 W(X_i).$$

Sarda (1993) considered such an estimator, but argued that “the resulting score function will produce a very small bandwidth.” Instead he introduced a “so-called cross-validation criterion”

$$CV(h) = n^{-1} \sum_{i=1}^n [\hat{F}_{h,-i}(X_i) - F_n(X_i)]^2 W(X_i),$$

where $\hat{F}_{h,-i}$ is the kernel estimator computed by leaving out X_i . In either case, the bandwidth minimizing the criterion is selected.

We now show that the criteria $CV(h)$ and $LNO(h)$ are asymptotically equivalent, under conditions (A.1)–(A.5) of Sarda (1993):

$$W \text{ is bounded and supported on a compact set.} \tag{A.1}$$

The set of bandwidths considered is

$$H_n = [C_1 n^{-a}, C_2 n^{-b}], \quad 1/4 < b \leq a < 1/2. \tag{A.2}$$

The function K is absolutely continuous and

$$\lim_{x \rightarrow -\infty} K(x) = 0 \text{ and } \lim_{x \rightarrow \infty} K(x) = 1. \tag{A.3}$$

For $k = K'$, we have

$$\int x k(x) dx = 0 \text{ and } \int x^2 k(x) dx < \infty \tag{A.4}$$

Also, F satisfies

F is twice differentiable and

F and $|f'|$ are bounded from below on the support of W . (A.5)

Under these conditions (weaker conditions are sufficient), it can be shown that

$$MISE(h) = V_1 n^{-1} - V_2 h n^{-1} + B_3 h^4 + C h^2/n + \text{smaller order terms}, \quad (2.1)$$

where $V_1 = D_1(F)$ and $V_2 = 2A_1(K)D_2(F)$ are the variance terms and $B_3 = 0.25[A_2(K)]^2 D_3(F)$ is the squared bias term and

$$A_1(K) = \int x k(x) K(x) dx \quad (2.2)$$

$$A_2(K) = \int x^2 k(x) dx \quad (2.3)$$

$$D_1(F) = \int F(x)(1 - F(x))f(x)W(x) dx \quad (2.4)$$

$$D_2(F) = \int [f(x)]^2 W(x) dx \quad (2.5)$$

$$D_3(F) = \int [f'(x)]^2 f(x) W(x) dx \quad (2.6)$$

and $C > 0$. So the asymptotically optimal bandwidth is

$$h_{opt} = (0.25V_2/B_3)^{1/3} n^{-1/3}. \quad (2.7)$$

Note that this is a second order correction since the leading term is $V_1 n^{-1}$ which is the same for all h , even $h = 0$ which corresponds to the EDF. Hence, let $d(\hat{F}_h, F)$ be either $MISE(h)$, $ASE(h)$ or $ISE(h)$ and let

$$d^*(\hat{F}_h, F) = d(\hat{F}_h, F) - V_1 n^{-1}.$$

We have the following theorem.

Theorem 1 *Under the assumptions (A.1)–(A.5), the leave-one-out criterion $CV(h)$ and the leave-none-out criterion $LNO(h)$ are asymptotically equivalent over H_n in the sense that*

$$\sup_{h \in H_n} \left| \frac{CV(h) - LNO(h)}{MISE^*(h)} \right| \rightarrow 0 \quad a.s.$$

Remark 1 Sarda (1993) showed that $CV(h)$ is asymptotically optimal with respect to d^* for choosing the bandwidth h in the sense that if \hat{h} minimizes $CV(h)$ over H_n then

$$\lim_{n \rightarrow \infty} \frac{d^*(\hat{F}_{\hat{h}}, F)}{\inf_{h \in H_n} d^*(\hat{F}_h, F)} = 1 \quad a.s.$$

An immediate consequence of Theorem 1 is that the bandwidth minimizing $LNO(h)$ is also asymptotically optimal with respect to d^* , *provided that Sarda's Theorem holds*. Theorems 2 and 3 and the simulation results seem incompatible with Sarda's Theorem.

Next we look at the expected value of $CV(h)$ and $LNO(h)$ and compare them with $MISE(h)$.

Theorem 2 *Assume that F has five derivatives and that the fifth derivative is bounded. Assume also that the kernel K has a symmetric density with respect to 0 and that its fifth moment exists. Finally, assume that the weight function has a finite first moment with respect to F . Then*

$$\begin{aligned} E[LNO(h)] &= 2\frac{h}{n}[A_3(K) - A_1(K)]D_2(F) + \frac{1}{4}h^4[A_2(K)]^2D_3(F) + O\left(\frac{1}{n^2}\right) + O\left(\frac{h^2}{n}\right) \\ E[CV(h)] &= 2\frac{h}{n}[A_3(K) - A_1(K)]D_2(F) + \frac{1}{4}h^4[A_2(K)]^2D_3(F) + O\left(\frac{1}{n^2}\right) + O\left(\frac{h^2}{n}\right) \end{aligned}$$

where $A_3(K) = \int_0^\infty vk(v) dv$, and $A_1(K)$ through $D_3(F)$ are defined in (2.2) through (2.6).

So the expected values of the two criteria are identical up to smaller order terms. Comparing with (2.1), note that neither criterion picks up the $O(1/n)$ term of $MISE(h)$; this is not important for the purpose of choosing the tuning parameter since this term does not depend on h . On the other hand, both criteria pick up an extra $O(h/n)$ term. This turns out to be very important since $A_3(K) - A_1(K) > \int_0^\infty vk(v)K(v) dv - \int_{-\infty}^\infty vk(v)K(v) dv > 0$. As a consequence of this extra term, the derivative of the expected value of $CV(h)$ or $LNO(h)$ is positive asymptotically for $h > 0$. The expected value of the derivative of $LNO(h)$ is also positive asymptotically as is shown in Theorem 3.

Theorem 3 *Suppose that the kernel k is symmetric about 0 and that the density has a continuous first derivative and a bounded second derivative on the support of the function W . Then the function $E[dLNO(h)/dh]$ is positive when $n \rightarrow \infty$ and $nh^2 \rightarrow \infty$.*

These two results suggest that neither *LNO* nor *CV* are likely to perform well in choosing the tuning parameter of a KDE, in sharp contradiction to the optimality theorem for *CV* of Sarda (1993). Simulations in Section 4 support this conclusion. The extra $O(h/n)$ term comes from the cross-product $E[\hat{F}_h(X_i)F_n(X_i)]$. Since the EDF uses all the observations, a covariance term of the form $D_1(F)A_3(K)$ appears from terms in X_j common to both the EDF and KDE. (The details are in the appendix). Using $\hat{F}_{h,-i}(X_i)$ instead of $F_h(X_i)$ is not sufficient since the remaining $n-1$ observations are common to both estimators. On the other hand, no such term is present if the estimators \hat{F}_h and F_n are based on two disjoint subsets.

For instance, let S be a subset of r elements from $\{1, 2, \dots, n\}$ with complement S^C and let

$$\hat{F}_h^S(x) = \frac{1}{r} \sum_{j \in S} K\left(\frac{x - X_j}{h}\right),$$

and

$$F_n^{S^C}(x) = \frac{1}{n-r} \sum_{j \in S^C} I(X_j \leq x).$$

Then for fixed x

$$\begin{aligned} E[(\hat{F}_h^S(x) - F_n^{S^C}(x))^2 W(x)] &= E[(\hat{F}_h^S(x) - F(x))^2 W(x)] + E[(F_n^{S^C}(x) - F(x))^2 W(x)] \\ &\quad + E[(\hat{F}_h^S(x) - F(x))(F(x) - F_n^{S^C}(x))W(x)] \\ &= E[(\hat{F}_h^S(x) - F(x))^2 W(x)] + E[(F_n^{S^C}(x) - F(x))^2 W(x)], \end{aligned} \quad (2.8)$$

since the two estimators are based on independent observations and $F_n^{S^C}(x)$ is unbiased for $F(x)$. The first term of (2.8) is the risk of \hat{F}_h based on r observations while the second term does not depend on h , and therefore does not play a role in the process of choosing the bandwidth. We do not pursue this idea further for a number of reasons. First, it is not sufficient that the expectation of the risk estimator be the right one to guarantee the convergence of the bandwidth of the minimizing estimate to the optimal bandwidth. For instance, while the cross-product has an expectation of 0, it is a random variable which, in some problems (e.g., Altman and Léger, 1994), may be just as large as the interesting term, $(\hat{F}_h(x) - F(x))^2 W(x)$, and therefore interferes in the bandwidth selection process. Second, this risk estimator is estimating the risk of \hat{F}_h based on $r < n$ observations. This is not a problem if r is close to n . But in that case, F_n is based on very few observations and will be highly variable. On the other hand, if $r \approx n - r$,

the variance of the two estimators will be balanced, but the selected bandwidth will be that for a sample of size r . To estimate the bandwidth for size n , a transformation based on an asymptotic development would need to be estimated. Instead, we introduce a plug-in estimator in the next section. Finally, note that the fundamental idea is not to leave out the observation X_i in the assessment of $\hat{F}_h(X_i)$, but rather to use independent subsets of observations for \hat{F}_h and F_n . If $r \approx n - r$, using X_i in the estimate of \hat{F}_h^S , in the estimate of F_n^{SC} or in neither leads to asymptotically equivalent formulas for $E(\hat{F}_h^S(X_i) - F_n^{SC}(X_i))^2 W(X_i)$.

Remark 2 The so-called cross-validation estimators of risk in distribution function and density estimation are based on two different heuristics. In distribution estimation, leaving some out reduces a covariance term between two different estimators of the true distribution function, one being the estimator of interest, the other one acting as a proxy for the unknown function.

In density estimation, the mean integrated squared error is

$$MISE(h) = E \int (\hat{f}_h(x) - f(x))^2 dx$$

No proxy for f is readily available. Instead the MISE is broken down as

$$MISE(h) = E \int \hat{f}_h(x)^2 dx - 2E \int \hat{f}_h(x)f(x) dx + \int f(x)^2 dx.$$

Noting that the last term is not a function of the bandwidth, estimators of the first two terms are introduced. The first term can be estimated by $\int \hat{f}_h(x)^2 dx$. For the second term, Stone (1984) noted that

$$\begin{aligned} E \int \hat{f}_h(x)f(x) dx &= Ek \left(\frac{X - Y}{h} \right) \\ &= E \frac{1}{n(n-1)} \sum_{i \neq j} k \left(\frac{X_i - X_j}{h} \right) \end{aligned}$$

where X and Y are independent random variables distributed according to the density f . Now $[n(n-1)]^{-1} \sum_{i \neq j} k((X_i - X_j)/h) = 1/n \sum \hat{f}_{h,-i}(X_i)$, where $\hat{f}_{h,-i}$ is the leave-one-out kernel density estimator of f . Using heuristics borrowed from a prediction context, Rudemo (1982) and Bowman (1984) have also introduced “cross-validation” estimates of $MISE(h) - \int f(x)^2 dx$, in which $\int \hat{f}_h(x)^2 dx$ is replaced by $1/n \sum \int \hat{f}_{h,-i}(x)^2 dx$. However, this criterion and Stone’s

criterion are asymptotically equivalent. So, the only place where it is crucial to use a leave-one-out kernel density estimator is in the estimate of the cross-product. The reason that it is crucial to leave out X_i is that $\hat{f}_{h;-i}(X_i) - \hat{f}_h(X_i) = (n-1)^{-1}\hat{f}_h(X_i) - [(n-1)h]^{-1}k(0)$ and the second term which is $O(1/nh)$ is as large as the MISE. Hence, leaving none out does not work in density estimation. On the other hand, $\hat{F}_{h;-i}(X_i) - \hat{F}_h(X_i) = (n-1)^{-1}\hat{F}_h(X_i) - (n-1)^{-1}K(0)$ and the second term does not depend on h . while the dependence of the first term on h is weak. Therefore, leaving out an observation in \hat{F}_h does not improve the criterion.

3 Plug-in Estimate of the Optimal Bandwidth h_{opt}

An alternative approach to selecting a bandwidth, is to use an estimator of the asymptotically optimal bandwidth. To estimate the asymptotically optimal bandwidth h_{opt} of equation (2.7), we must first estimate V_2 and B_3 . As suggested by Hall and Marron (1987), we use the following kernel estimator of V_2 :

$$\begin{aligned}\hat{V}_2 &= 2A_1(K)\hat{D}_2(F) \\ &= 2A_1(K) \times \frac{1}{n(n-1)} \sum_{i \neq j} \alpha_v^{-1} k_v \left(\frac{X_i - X_j}{\alpha_v} \right) W(X_i)\end{aligned}$$

where k_v is a kernel, α_v is its bandwidth, K is the kernel used in computing the KDE and $A_1(K)$ is defined in (2.2). Using assumption (A.5) and Theorem 3.3(b) of Hall and Marron (1987), \hat{V}_2 is consistent when α_v satisfies $n\alpha_v \rightarrow \infty$ and $n^{1/4}\alpha_v \rightarrow 0$.

To estimate $D_3(F)$, we introduce the following kernel estimator:

$$\hat{D}_3(F) = \frac{1}{n^3\alpha_b^4} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n k'_b \left(\frac{X_i - X_j}{\alpha_b} \right) k'_b \left(\frac{X_i - X_k}{\alpha_b} \right) W(X_i), \quad (3.1)$$

where k'_b is the derivative of a kernel k_b and α_b is the associated bandwidth.

The estimator of B_3 is $0.25\hat{D}_3(F)[A_2(K)]^2$, leading to

$$\hat{h}_{opt} = (0.25\hat{V}_2/\hat{B}_3)^{1/3}n^{-1/3} \quad (3.2)$$

as the plug-in estimator of h_{opt} . The bias and variance of $\hat{D}_3(F)$ are computed in the following Theorem.

Theorem 4 *Suppose that the kernel k_b has mean 0, finite variance and that $k'_b(0) = 0$. Suppose also that the density f has a bounded fourth derivative. Then as $n \rightarrow \infty$ and $\alpha_b \rightarrow 0$ with $n\alpha_b \rightarrow \infty$,*

$$\begin{aligned} (a) \text{ Bias}(\hat{D}_3(F)) &= -\alpha_b^2 \int f(x) f'(x) f'''(x) W(x) dx \int u^2 k_b(u) du \\ &\quad - n^{-2} \alpha_b^{-3} \int (f(x))^2 W(x) dx \int (k'_b(u))^2 du + O(1/n) \\ (b) \text{ Var}(\hat{D}_3(F)) &= 2n^{-2} \alpha_b^{-5} \int (f(x))^4 W(x) dx \times C(k_b) + O(1/n), \\ \text{where } C(k_b) &= \int \int \int k'_b(s) k'_b(t) k'_b(u) k'_b(t+u-s) ds dt du. \end{aligned}$$

Note that the condition $k'_b(0) = 0$ is not necessary but simplifies computations as terms in (3.1) are zero when $i = j$ or $i = k$.

The asymptotically optimal choice of bandwidth is immediate.

Remark 3 *Under the conditions of Theorem 4, the best rates of convergence of the mean squared errors are as follows:*

- (a) *If $C(k_b) \neq 0$, $E(\hat{D}_3(F) - D_3(F))^2 = O(n^{-8/9})$ provided that $\alpha_b = cn^{-2/9}$ for some fixed constant c ;*
- (b) *If $C(k_b) = 0$, $E(\hat{D}_3(F) - D_3(F))^2 = O(n^{-1})$, provided that $n\alpha_b \rightarrow \infty$ and $n\alpha_b^4 \rightarrow 0$.*

Remark 4 For general kernels k , $C(k) \neq 0$. However, for the quadratic kernel $k(x) = 3/4(1 - x^2)$, $-1 \leq x \leq 1$, $C(k) = 0$. Therefore, using this kernel improves the rate of convergence and drastically simplifies the choice of bandwidth α_b .

4 Simulations

A small simulation study was run to assess the small sample behavior of the three bandwidth selectors: the minimizers of $CV(h)$ and $LNO(h)$, and the plug-in estimator \hat{h}_{opt} of (3.2). The quadratic kernel defined by $k(x) = \frac{3}{4}(1 - x^2)$ for $-1 \leq x \leq 1$ was used to compute the KDE.

An experiment involves a distribution for the observations, a weight function $W(x)$, and a sample size. We used samples from five distributions. The exponential distribution with $W(x) = I([.1, 3])$, where $I(A)$ is the indicator function of the set A , was used to satisfy the conditions of our Theorems, including conditions (A.5). Four target distributions were also

| Name | Distribution Function |
|-----------------------------------|--|
| Standard Exponential Distribution | $\text{Exp}(1)$ |
| Standard Normal Distribution | $N(0, 1)$ |
| Skewed Unimodal Distribution #2 | $\frac{1}{5}N(0, 1) + \frac{1}{5}N(\frac{1}{2}, \frac{4}{9}) + \frac{3}{5}N(\frac{13}{12}, \frac{25}{81})$ |
| Skewed Bimodal Distribution #8 | $\frac{3}{4}N(0, 1) + \frac{1}{4}N(\frac{3}{2}, \frac{1}{9})$ |
| Claw Distribution #10 | $\frac{1}{2}N(0, 1) + \sum_{i=0}^4 \frac{1}{10}N(\frac{i}{2} - 1, \frac{1}{100})$ |

Table 1: Distribution functions used in the simulation study. Plots of the Normal mixture densities are in Marron and Wand, 1992.

selected from the examples of normal mixtures given in Marron and Wand (1992). These were: standard normal distribution, skewed unimodal distribution #2, asymmetric bimodal distribution #8 and claw distribution #10. The cumulative distribution functions are listed in Table 1. For these four distributions, the weight function was $W(x) = I([-3, 3])$. For each of these five cases, 500 samples of sizes 10, 100 and 1000 were generated. For each sample, the minimizer of $CV(h)$, $LNO(h)$ and $ASE(h)$ was determined over a logarithmic grid of 100 bandwidths from .004 to 3.73. For samples of size 10 and 100, the plug-in estimator \hat{h}_{opt} was computed. As the plug-in estimator requires $O(n^3)$ computations (to evaluate \hat{B}_3) we did not include it for samples of size 1000. Along with the bandwidth estimates, we computed $ASE(\hat{h})$ where \hat{h} is the bandwidth estimate. For each fixed bandwidth h , we also computed the average (over the 500 samples) of $CV(h)$, $LNO(h)$, and $ASE(h)$.

The plug-in estimator requires two kernels to estimate V_2 and B_3 . Following Remark 4, in both cases we use the quadratic kernel with $\alpha = n^{-.3}$ as a pilot bandwidth.

Remark 5 Note that for the mixtures of normals, the combinations of distribution and weight function do not satisfy condition (A.5). We are not convinced that condition (A.5) is necessary. Sarda's Theorem was proven using a development inspired by Härdle and Marron (1985). They needed a similar condition because their kernel regression estimator had a kernel density estimator in its denominator. Sarda's estimator does not have such a random denominator. Moreover, if condition (A.5) is necessary, then the weight function must be 0 in the neighbor-

hood of a mode, a very strict limitation on the usefulness of the results.

We present the results for the exponential distribution as they are representative of the other cases. Figure 1 shows boxplots of the selected bandwidths by the four criteria, for each of the three sample sizes. Note that the bandwidth minimizing ASE is not available in practice since it requires knowing the true distribution function. The horizontal full line is the bandwidth minimizing the expected ASE whereas the broken line is h_{opt} which minimizes the asymptotic $MISE$. For samples of size 10, the bandwidths selected by CV and LNO are always smaller than the bandwidth minimizing the expected ASE and are smaller than the first quartile of the (random) bandwidth minimizing ASE . On the other hand, the plug-in bandwidth does very well. For samples of size 100 and 1000, CV and LNO do terribly: in the 500 simulated samples, they always picked the smallest available bandwidth, $h = .04$. For samples of size 100, the plug-in estimator continues to do well.

A weakness of plug-in estimators is that they estimate the asymptotically optimal bandwidth, which may not be close to the optimal bandwidth for a fixed sample size. Nevertheless, there is no doubt that the plug-in estimator is doing much better than the other two methods.

Next, we empirically try to validate Sarda's Theorem about the optimality of the CV selected distribution estimator. From Remark 1, if CV is optimal, so is LNO . Let h_{ASE} minimize $ASE(h)$ over the grid of 100 bandwidths and let \hat{h} be the selected bandwidth of a given method. We compute

$$\left| \frac{ASE(\hat{h}) - V_1/n}{ASE(h_{ASE}) - V_1/n} - 1 \right|. \quad (4.1)$$

This quantity should converge to 0 almost surely for the minimizers of CV and of LNO . Recall that V_1/n is subtracted because this term is common to all KDEs, including the EDF ($h = 0$). The random variable (4.1) was computed for the different methods for each of the 500 samples for each sample size. Figure 3 shows boxplots; a number of large outliers have been hidden to emphasis the detail of the main section of the boxes. Of the three methods, the plug-in is more concentrated about 0 than the other two, even for a sample size as small as 10. Even for samples as large as 1000, CV and LNO are less concentrated about 0 than the plug-in for 10 observations. These empirical statistics do not seem to confirm the asymptotic result of Sarda.

Finally, in Figure 2 we look at the expected value of the CV , LNO , and ASE criteria over

the 500 simulations for each fixed bandwidth h . As suggested by Theorem 3, the expected value of $CV(h)$ and $LNO(h)$ are increasing functions of h . When the sample size is small, the variation about the expected value is such that the selected bandwidth is not necessarily the smallest available bandwidth. But as n increases, the variation diminishes and the smallest bandwidth is almost always selected. On the other hand, there is a clearly visible minimum of the expected ASE . Note that the ratio of the expected ASE for h close to 0 over the smallest expected ASE is not very large. This is consistent with the numbers obtained in Table 2.

The simulation results clearly show that neither $CV(h)$ nor $LNO(h)$ are sufficiently good for bandwidth selection in the range of sample sizes observed in practice. They also shed considerable doubts on the optimality result of Sarda (1993) for the minimizer of CV . Finally, the plug-in estimator introduced in the previous section does well.

5 Discussion

Leave-one-out cross-validation for choosing among competing estimators was first suggested by Stone (1974) as an estimator of average *prediction* risk. Later Clark (1975) and Wahba and Wold (1975) applied it to selecting the smoothing parameter for smoothing splines. Leaving an observation out of the estimator to predict that observation mimics the prediction problem. If estimation risk rather than prediction risk is of interest, prediction based estimates of risk can still be used if the L_2 loss is used since the prediction risk is the estimation risk plus a constant independent of the smoothing parameter.

The suggestion of using a similar criterion in density estimation is due to Bowman (1984) and Rudemo (1982). However, although Bowman introduced a heuristic linking least squares cross-validation for density estimation to a prediction problem, prediction risk is not well defined for density estimation. Leaving one out in that problem is actually a bias reduction technique for the estimation risk estimator. Leave-one-out bias reduction techniques require some care in application. As we have shown here, results do not readily generalize from problem to problem. In kernel density estimation, use of the leave-one-out estimator in the selection criterion introduces extra terms depending on the bandwidth and of the same order of magnitude as the squared error risk; these are removed by use of a leave-one-out estimator. In KDE, use of

the leave-none-out estimator in the selection criterion also introduces terms depending on the bandwidth which are the same order of magnitude as the squared error risk; however, use of the leave-one-out KDE does not remove these terms. Instead, the kernel estimator and the proxy for the unknown function must be based on disjoint subsets of independent observations, as was done by Müller, Stadtmüller, and Schmitt (1987) for the nonparametric regression derivative problem.

Another interesting case where the leave-some-out estimators were used explicitly for bias reduction is in the estimator of integrated squared density derivatives:

$$\int f^{(p)}(x)^2 dx$$

proposed by Hall and Marron (1987). They noted that

$$\int \hat{f}_h^{(p)}(x)^2 dx = \frac{1}{n^2 h^{2p+1}} \sum_{i=1}^n \sum_{j=1}^n K^{(p)} * K^{(p)} \left(\frac{y_i - y_j}{h} \right)$$

where “*” denotes convolution. They argue that the terms for which $i = j$ do not involve the data and can be thought of as bias terms, and so they propose an estimator which explicitly excludes those terms. However, Sheather and Jones (1991) showed that the excluded terms can actually be used to improve the estimator by cancelling other bias terms.

Cross-validation and leave-some-out estimators are useful tools in the statistician’s toolkit, but care needs to be taken in application. Whether the motivation for their use is prediction risk estimation or bias reduction, it is necessary to check both analytically and by simulation whether the behavior of the selection criterion is similar to that of the risk function one wishes to minimize. Some of the known problems with the selection criteria in various contexts include excessive variability, bias of the same order as the risk function, and mean zero stochastic terms of same order in probability as the loss function.

Plug-in methods appear to have first been suggested by Woodroofe (1970) but technical difficulties were not overcome until recently. Recent results in both kernel regression (Gasser, Kneip, and Köhler, 1991) and density estimation (Jones, Marron and Park, 1991; Park and Marron, 1991; Sheather and Jones, 1991) show that plug-in approaches can come close to the optimal relative rate of convergence for estimating the asymptotically optimal bandwidth. Preliminary results presented in this note demonstrate that plug-in bandwidths perform well for

kernel distribution function estimation as well, although no rate results are currently available. One problem to be overcome, however, is that although plug-in bandwidths are good estimators of the asymptotically optimal bandwidth, the optimal bandwidth at small sample sizes may be somewhat larger.

6 Appendix

Proof of Theorem 1: After using algebra similar to the proof of Lemma 2 of Sarda (1993), we have

$$\begin{aligned}
CV(h) - LNO(h) &= n^{-1} \sum_{i=1}^n [\hat{F}_{h;-i}(X_i) - \hat{F}_h(X_i)]^2 W(X_i) + \\
&\quad 2n^{-1} \sum_{i=1}^n (\hat{F}_{h;-i}(X_i) - \hat{F}_h(X_i))(\hat{F}_h(X_i) - F_n(X_i))W(X_i) \\
&= n^{-1} \sum_{i=1}^n A_i^2 W(X_i) + 2n^{-1} \sum_{i=1}^n A_i(\hat{F}_h(X_i) - F_n(X_i))W(X_i) \\
&= (\overline{ASE}(h) - ASE(h)) + 2n^{-1} \sum_{i=1}^n A_i(F(X_i) - F_n(X_i))W(X_i) \\
&= T_1(h) + T_2(h),
\end{aligned}$$

where, as in Sarda, $A_i = (n-1)^{-1}(\hat{F}_h(X_i) - K(0))$ and $\overline{ASE}(h) = n^{-1} \sum_{i=1}^n [\hat{F}_{h;-i}(X_i) - F(X_i)]^2 W(X_i)$.

From the same proof,

$$\sup_{h \in \tilde{H}_n} \left| \frac{T_1(h)}{MISE^*(h)} \right| \rightarrow 0, \quad a.s.$$

Also, using inequality (3.2) of Sarda, assumption (A.2), and for $a < 1/2$, we have

$$\begin{aligned}
\sup_{h \in \tilde{H}_n} \left| \frac{T_2(h)}{MISE^*(h)} \right| &\leq Cn^{1+a} \sup_{h \in \tilde{H}_n} |T_2(h)| \\
&\leq 4C^*n^a \sup_x |F_n(x) - F(x)|,
\end{aligned}$$

since W is bounded from (A.1) and $|A_i| \leq 2/(n-1)$. Now, Smirnov's law of the iterated logarithm (Shorack and Wellner, 1986, p. 504) states that

$$\lim_{n \rightarrow \infty} \frac{\sqrt{n} \sup_x |F_n(x) - F(x)|}{\sqrt{2 \log_2 n}} = 1/2 \quad a.s.$$

Hence since $a < 1/2$, we have that

$$\sup_{h \in H_n} \left| \frac{T_2(h)}{MISE^*(h)} \right| \rightarrow 0 \quad a.s.$$

Proof of Theorem 2: We note that

$$\begin{aligned} E(LNO(h)) &= E \left[\left(\hat{F}_h[X_i] - F_n(X_i) \right)^2 W(X_i) \right] \\ &= \frac{1}{n^2} E \left[\left(\frac{1}{2} + \sum_{i \neq j} K \left(\frac{X_i - X_j}{h} \right) - 1 - \sum_{i \neq j} I(X_j \leq X_i) \right)^2 W(X_i) \right] \\ &= \frac{1}{n^2} E \left[\frac{W(X_i)}{4} + (n-1) \left(K \left(\frac{X_i - X_j}{h} \right)^2 W(X_i) - K \left(\frac{X_i - X_j}{h} \right) W(X_i) \right) \right. \\ &\quad + 2(n-1) \left(I(X_j \leq X_i) W(X_i) - K \left(\frac{X_i - X_j}{h} \right) I(X_j \leq X_i) W(X_i) \right) \\ &\quad + (n-1)(n-2) \left(K \left(\frac{X_i - X_j}{h} \right) K \left(\frac{X_i - X_k}{h} \right) W(X_i) \right. \\ &\quad \left. \left. + I(X_j \leq X_i) I(X_k \leq X_i) W(X_i) \right) \right. \\ &\quad \left. - 2(n-1)(n-2) K \left(\frac{X_i - X_j}{h} \right) I(X_k \leq X_i) W(X_i) \right] \end{aligned}$$

and

$$\begin{aligned} E(CV(h)) &= E(LNO(h)) + \frac{1}{(n-1)^2} E \left(\hat{F}_h(X_i) - \frac{1}{2} \right)^2 W(X_i) \\ &\quad + \frac{2}{n-1} E \left(\hat{F}_h(X_i) - \frac{1}{2} \right) \left(\hat{F}_h(X_i) - F_n(X_i) \right) W(X_i) \\ &= E(LNO(h)) + \frac{1}{n^2} E \left[-\frac{3W(X_i)}{4} + (n-1) K \left(\frac{X_i - X_j}{h} \right) W(X_i) \right. \\ &\quad + \frac{2n-1}{n-1} K \left(\frac{X_i - X_j}{h} \right)^2 W(X_i) + \frac{(2n-1)(n-2)}{n-1} K \left(\frac{X_i - X_j}{h} \right) K \left(\frac{X_i - X_k}{h} \right) W(X_i) \\ &\quad + (n-1) I(X_j \leq X_i) W(X_i) - 2K \left(\frac{X_i - X_j}{h} \right) I(X_j \leq X_i) W(X_i) \\ &\quad \left. - 2(n-2) K \left(\frac{X_i - X_j}{h} \right) I(X_k \leq X_i) W(X_i) \right] \end{aligned}$$

We need to evaluate the leading terms of the expectations. By changing the order of

integration and expanding $F(t - hv)$ into a Taylor series around t we obtain:

$$E \left[K \left(\frac{X_i - X_j}{h} \right) W(X_i) \right] = D_4(F) + O(h^2) \quad (6.1)$$

$$E[I(X_j \leq X_i)W(X_i)] = D_4(F) \quad (6.2)$$

$$E \left[K \left(\frac{X_i - X_j}{h} \right) I(X_j \leq X_i)W(X_i) \right] = D_4(F) - hA_3(K)D_2(F) + O(h^2) \quad (6.3)$$

$$\begin{aligned} E \left[K \left(\frac{X_i - X_j}{h} \right) I(X_k \leq X_i)W(X_i) \right] &= D_5(F) + \frac{h^2}{2} A_2(K)D_6(F) \\ &\quad + \frac{h^4}{24} A_4(K)D_7(F) + O(h^6) \end{aligned} \quad (6.4)$$

$$E \left[K \left(\frac{X_i - X_j}{h} \right)^2 W(X_i) \right] = D_4(F) - 2hA_1(K)D_2(F) + O(h^2) \quad (6.5)$$

$$\begin{aligned} E \left[K \left(\frac{X_i - X_j}{h} \right) K \left(\frac{X_i - X_k}{h} \right) W(X_i) \right] &= D_5(F) + h^2 A_2(K)D_6(F) + \frac{h^4}{12} A_4(K)D_7(F) \\ &\quad + \frac{1}{4} h^4 [A_2(K)]^2 D_3(F) + O(h^6) \end{aligned} \quad (6.6)$$

$$E[I(X_j \leq X_i)I(X_k \leq X_i)W(X_i)] = D_5(F), \quad (6.7)$$

where $D_4(F) = \int F(x)f(x)W(x)dx$, $D_5(F) = \int f(x)F^2(x)W(x)dx$, $D_6(F) = \int f'(x)F(x)f(x)W(x)dx$, $D_7(F) = \int f^{(3)}(x)F(x)f(x)W(x)dx$, and $A_4(K) = \int x^4 k(x)dx$.

The result then follows by substitution.

Proof of Theorem 3: The derivative of $LNO(h)$ with respect to h is

$$\frac{d LNO(h)}{dh} = \frac{2}{n} \sum_{i=1}^n [\hat{F}_h(X_i) - F_n(X_i)] W(X_i) \frac{d \hat{F}_h(X_i)}{dh}.$$

But

$$\frac{d \hat{F}_h(x)}{dh} = \frac{d}{dh} \frac{1}{n} \sum_{i=1}^n K \left(\frac{x - X_i}{h} \right) = \frac{1}{nh} \sum_{i=1}^n k \left(\frac{x - X_i}{h} \right) \left(\frac{X_i - x}{h} \right)$$

Hence,

$$\begin{aligned} \frac{d LNO(h)}{dh} &= \frac{2}{n} \sum_{i=1}^n \left[\left(\hat{F}_h(X_i) - F_n(X_i) \right) \left(\frac{1}{nh} \sum_{j=1}^n k \left(\frac{X_i - X_j}{h} \right) \left(\frac{X_j - X_i}{h} \right) \right) W(X_i) \right] \\ &= \frac{2}{n^3 h} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \left\{ \left[K \left(\frac{X_i - X_j}{h} \right) - I(X_i - X_j \geq 0) \right] k \left(\frac{X_i - X_k}{h} \right) \left(\frac{X_k - X_i}{h} \right) W(X_i) \right\}, \end{aligned}$$

where $I(\cdot)$ is the indicator function.

We must now take the expectation of this expression. Consider first the case where i, j and k are all different.

$$\begin{aligned}
E \left\{ \frac{1}{h} \left[K \left(\frac{X_i - X_j}{h} \right) - I(X_i - X_j \geq 0) \right] k \left(\frac{X_i - X_k}{h} \right) \left(\frac{X_k - X_i}{h} \right) W(X_i) \right\} \\
= \int \int \int \left[K \left(\frac{x - y}{h} \right) - I(x - y \geq 0) \right] \frac{1}{h} k \left(\frac{x - z}{h} \right) \left(\frac{z - x}{h} \right) W(x) f(x) f(y) f(z) dx dy dz \\
= \int \int \int [I(x \geq 0) - K(x)] k(y) y h W(z) f(z) f(z - xh) f(z - yh) dx dy dz \\
= A_2(K) A_5(K) D_3(F) h^3 + o(h^3),
\end{aligned}$$

by a Taylor series expansion of f where $A_5(K) = \int [I(x \geq 0) - K(x)] x dx > 0$, by the symmetry of k . Note that $A_2(K)$ and $D_3(F)$, defined in (2.3) and (2.6), are both positive.

When $i = j \neq k$, we have

$$\begin{aligned}
E \left\{ \frac{1}{h} [K(0) - 1] k \left(\frac{X_i - X_k}{h} \right) \left(\frac{X_k - X_i}{h} \right) W(X_i) \right\} \\
+ 1/2 \int \int \frac{1}{h} k \left(\frac{x - y}{h} \right) \left(\frac{y - x}{h} \right) W(x) f(x) f(y) dx dy \\
= 1/2 \int \int u k(u) W(v) f(v) f(v - uh) du dv \\
= -1/2 A_2(K) D_2(F) h + o(h).
\end{aligned}$$

Finally, when $i = k$, the random variable is 0. Thus,

$$\begin{aligned}
E(LNO(h)) &= 2A_2(K) A_5(K) D_3(F) h^3 - A_2(K) D_2(F) h n^{-1} + o(h^3 + h n^{-1}) \\
&> 0,
\end{aligned}$$

provided that n is sufficiently large and that $nh^2 \rightarrow \infty$.

Proof of Theorem 4 and Remark 3: Note that after a change of variables

$$\begin{aligned}
E \left[\frac{1}{n^3 \alpha_b^4} k'_b \left(\frac{X_i - X_j}{\alpha_b} \right) k'_b \left(\frac{X_i - X_k}{\alpha_b} \right) W(X_i) \right] \\
= \frac{1}{n^3} \int \int \int k_b(s) k_b(t) f(u) f'(u - s\alpha_b) f'(u - t\alpha_b) W(u) ds dt du. \quad (6.8)
\end{aligned}$$

After a second order Taylor series expansion of $f'(u - s\alpha_b)$ and $f'(u - t\alpha_b)$, we obtain the first bias term. The other bias term comes from

$$E \left\{ \frac{1}{n^3 \alpha_b^4} \left[k'_b \left(\frac{X_i - X_j}{\alpha_b} \right) \right]^2 W(X_i) \right\} = \frac{1}{n^3 \alpha_b^3} \int \int [k'_b(s)]^2 f(t) f(t - s\alpha_b) W(t) ds dt.$$

A first order Taylor series expansion of $f'(t - s\alpha_b)$ gives the second bias term.

The variance is given by

$$\begin{aligned} \text{Var}(\hat{D}_3(F)) &= \frac{1}{n^6 \alpha_b^8} \times \\ &\sum_{i,j,k,i',j',k'} \text{Cov} \left[k'_b \left(\frac{X_i - X_j}{\alpha_b} \right) k'_b \left(\frac{X_i - X_k}{\alpha_b} \right) W(X_i), k'_b \left(\frac{X_{i'} - X_{j'}}{\alpha_b} \right) k'_b \left(\frac{X_{i'} - X_{k'}}{\alpha_b} \right) W(X_{i'}) \right]. \end{aligned} \quad (6.9)$$

This is very tedious to compute, but the leading term is when i, j, k and i' are all different, and $\{j, k\} = \{j', k'\}$. We have

$$\begin{aligned} &\frac{1}{n^6 \alpha_b^8} E \left[k'_b \left(\frac{X_i - X_j}{\alpha_b} \right) k'_b \left(\frac{X_i - X_k}{\alpha_b} \right) W(X_i) k'_b \left(\frac{X_{i'} - X_j}{\alpha_b} \right) k'_b \left(\frac{X_{i'} - X_k}{\alpha_b} \right) W(X_{i'}) \right] \\ &= \frac{1}{n^6 \alpha_b^5} \int \int \int \int k'_b(s) k'_b(t) k'_b(u) k'_b(t+u-s) \\ &\quad \times f(v) f(v - s\alpha_b) f(v - t\alpha_b) f(v - s\alpha_b + u\alpha_b) W(v) W(v - s\alpha_b + u\alpha_b) ds dt du dv. \end{aligned} \quad (6.10)$$

After a Taylor series expansion and since there are $O(n^4)$ such terms, we obtain the leading term of the variance. Note from (6.8) that $E \left[\frac{1}{n^3 \alpha_b^4} k'_b \left(\frac{X_1 - X_2}{\alpha_b} \right) k'_b \left(\frac{X_1 - X_3}{\alpha_b} \right) W(X_1) \right] = O(n^{-3})$, so that the product of the expectations in the covariance term are small compared to the expectation of the product in (6.10).

The remaining covariance terms are $O(n^{-1})$, $O(n^{-3} \alpha_b^{-6})$, $O(n^{-3} \alpha_b^{-5})$, $O(n^{-2} \alpha_b^{-3})$, and $O(n^{-4} \alpha_b^{-7})$. To balance squared bias and variance, we must balance $O(\alpha_b^4)$ and $O(n^{-2} \alpha_b^{-5})$, leading to $\alpha_b = O(n^{-2/9})$ and an expected squared error of $O(n^{-8/9})$.

Note that if $C(k_b) = \int \int \int k'_b(u) k'_b(v) k'_b(w) k'_b(v+w-u) du dv dw = 0$, then the first non-zero term in (6.10) is the term corresponding to

$$\frac{1}{n^6 \alpha_b^3} \int \int \int \int s t k'_b(s) k'_b(t) k'_b(u) k'_b(t+u-s) [f(v)]^2 [f'(v)]^2 W^2(v) ds dt du dv,$$

so that these covariance terms are $O(n^{-2} \alpha_b^{-3})$. In this case the leading variance term is $O(n^{-1})$ provided that $n \alpha_b^4 \rightarrow 0$.

References

- Altman, N. and Léger, C. (1993). "Cross-validation, the bootstrap, and related methods for tuning parameter selection." Technical Report BU-1216-M, Biometrics Unit, Cornell University.
- Bowman, A. W. (1984) "An Alternative Method of Cross-validation for the smoothing of density estimates," *Biometrika*, **71**, 353-360.
- Clark, R. M., (1975) "A Calibration Curve for Radiocarbon Dates," *Antiquity*, **49** 251-266.
- Gasser, T., Kneip, A., Köhler, W. (1991) "A Flexible and Fast Method for Automatic Smoothing," *Journal of the American Statistical Association*, **86**, 643-652.
- Hall and Marron (1987) "Estimation of Integrated Squared Density Derivatives," *Statistics and Probability Letters* **6**, 109-115.
- Härdle, W., and Marron, J. S. (1985) "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *Annals of Statistics*, **13**, 1465-1481.
- Jones, M. C., Marron, J. S. and Park, B. U. (1991) "A Simple Root n Bandwidth Selector," *Annals of Statistics* **19**, 1919-1932.
- Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Annals of Statistics* **20** 712-736.
- Marron, J. S. and Wand, M. P. (1992). "Exact mean integrated squared error." *Annals of Statistics* **20** 712-736.
- Müller, H.-G., Stadtmüller, U. and Schmitt, T. (1987) "Bandwidth Choice and Confidence Intervals for Derivatives of Noisy Data," *Biometrika* **74**, 743-749.
- Nadaraya, E. A. (1964) "On estimating regression," *Theory of Probability and its Applications*, **9**, 141-142.
- Park, B. U. and Marron, J. S. (1990) "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistical Association*, **85**, 66-72.
- Rudemo, M. (1982) "Empirical Choice of Histograms and Kernel Density Estimators," *Scandinavian Journal of Statistics*, **9**, 65-78.
- Sarda, P. (1993). "Smoothing parameter selection for smooth distribution functions." *Journal of Statistical Planning and Inference* **35** 65-75.

- Sheather, S. J. and Jones, M. C. (1991) "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society B* **53**, 683-690.
- Stone, C. J. (1977) "Consistent Nonparametric Regression," *Annals of Statistics*, **5**, 595-645.
- Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions." *Journal of the Royal Statistical Society B* **36** 111-147.
- Stone, C. J. (1984). "An asymptotically optimal window selection rule for kernel density estimates." *Annals of Statistics* **12** 1285-1297.
- Wahba and Wold (1975) "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation," *Communications in Statistics*, **4**, 1-47.
- Woodroffe, M. (1970) "On Choosing a Delta Sequence," *The Annals of Mathematical Statistics*, **41**, 1665-1671.

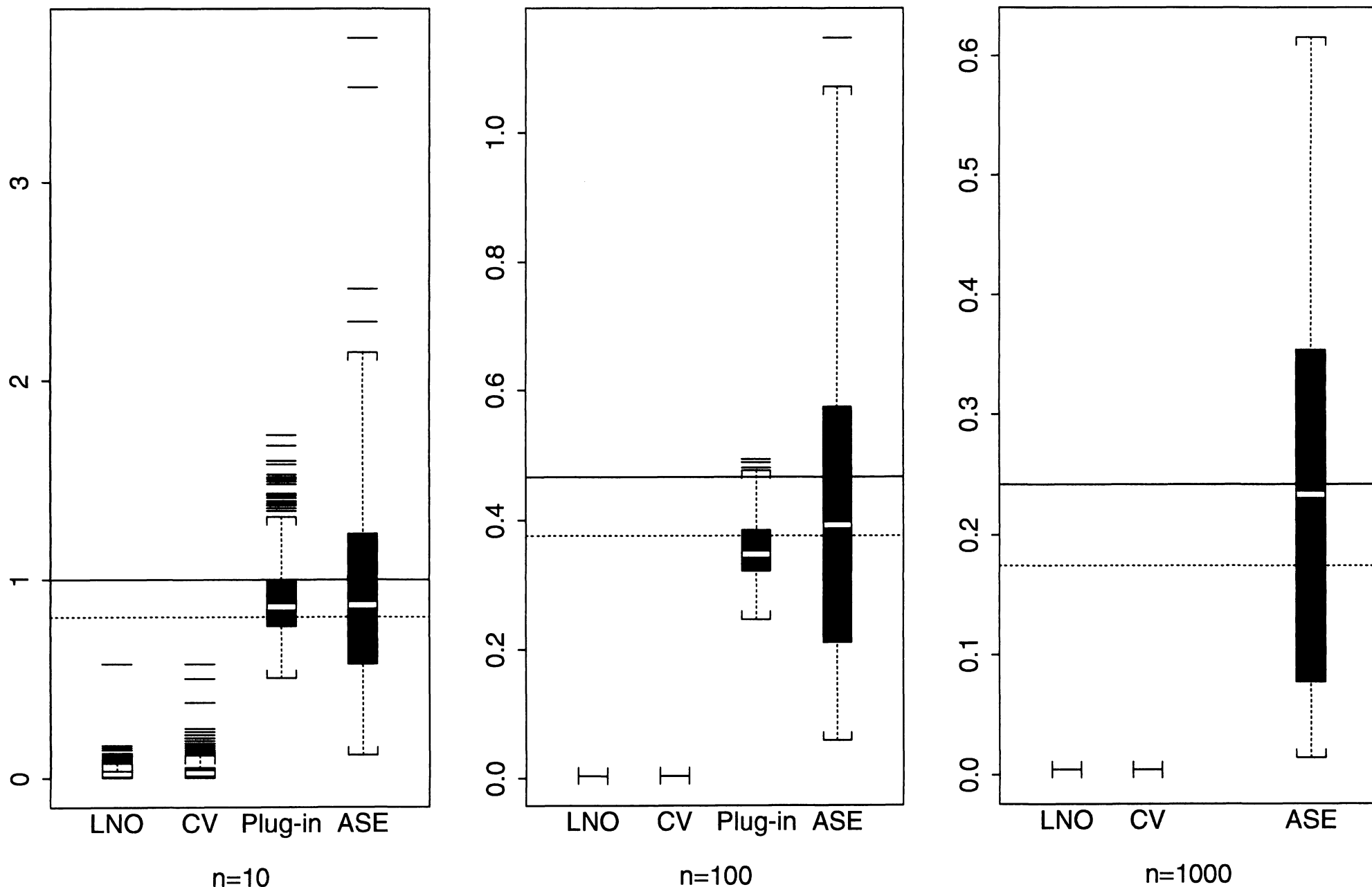


FIGURE 1: The bandwidths minimizing the LNO, CV, Plug-in, and ASE criteria for 500 samples of size n from an exponential distribution. The horizontal full and dashed lines are the bandwidth minimizing the expected ASE and the asymptotic MISE, respectively. The weight function is the indicator of the interval $[.1,3]$.

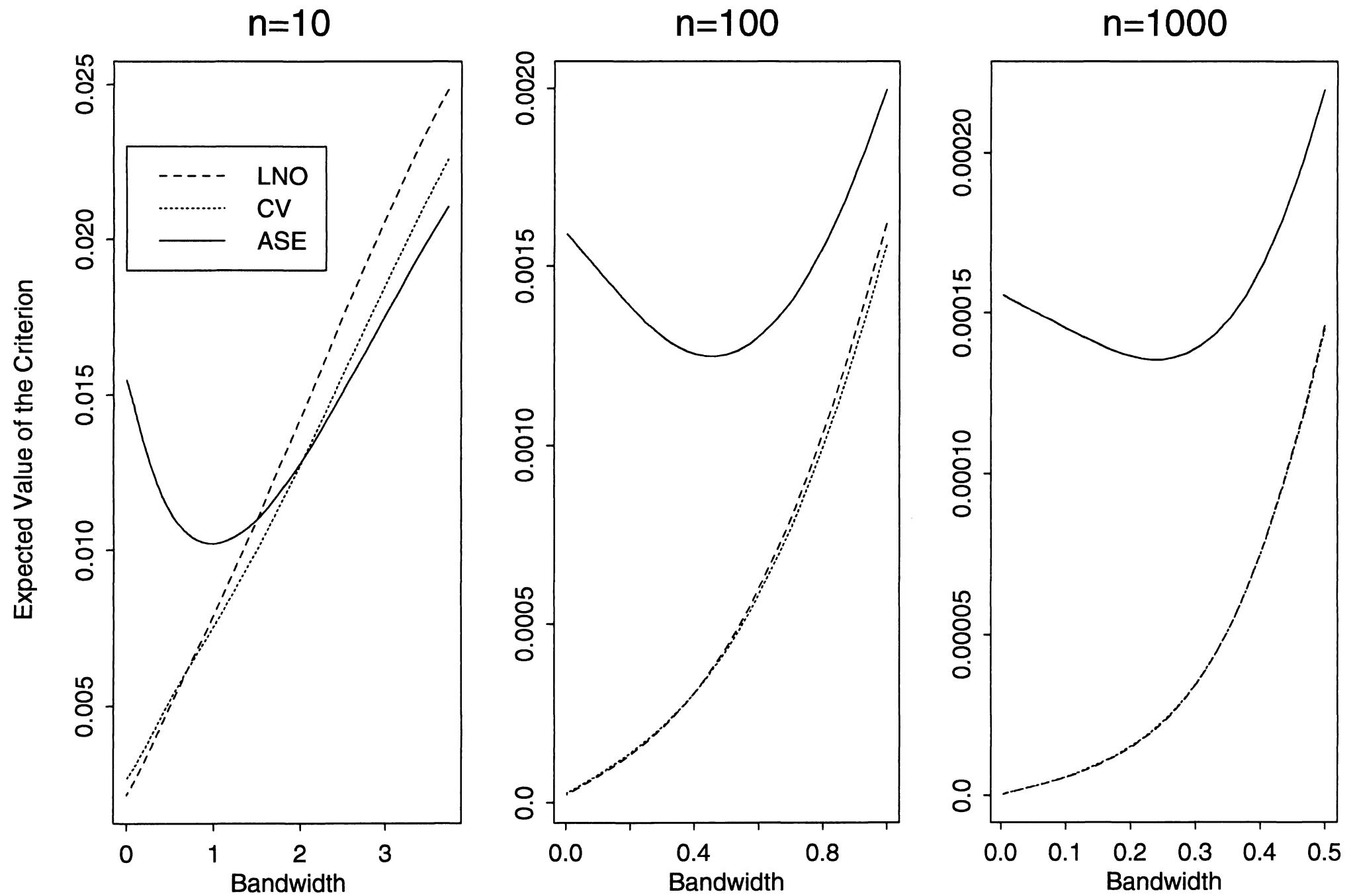


FIGURE 2: The expected value of CV, LNO, and ASE for each fixed bandwidth over 500 simulations of samples of size n from the exponential distribution. The weight function is the indicator of the interval $[.1,3]$.

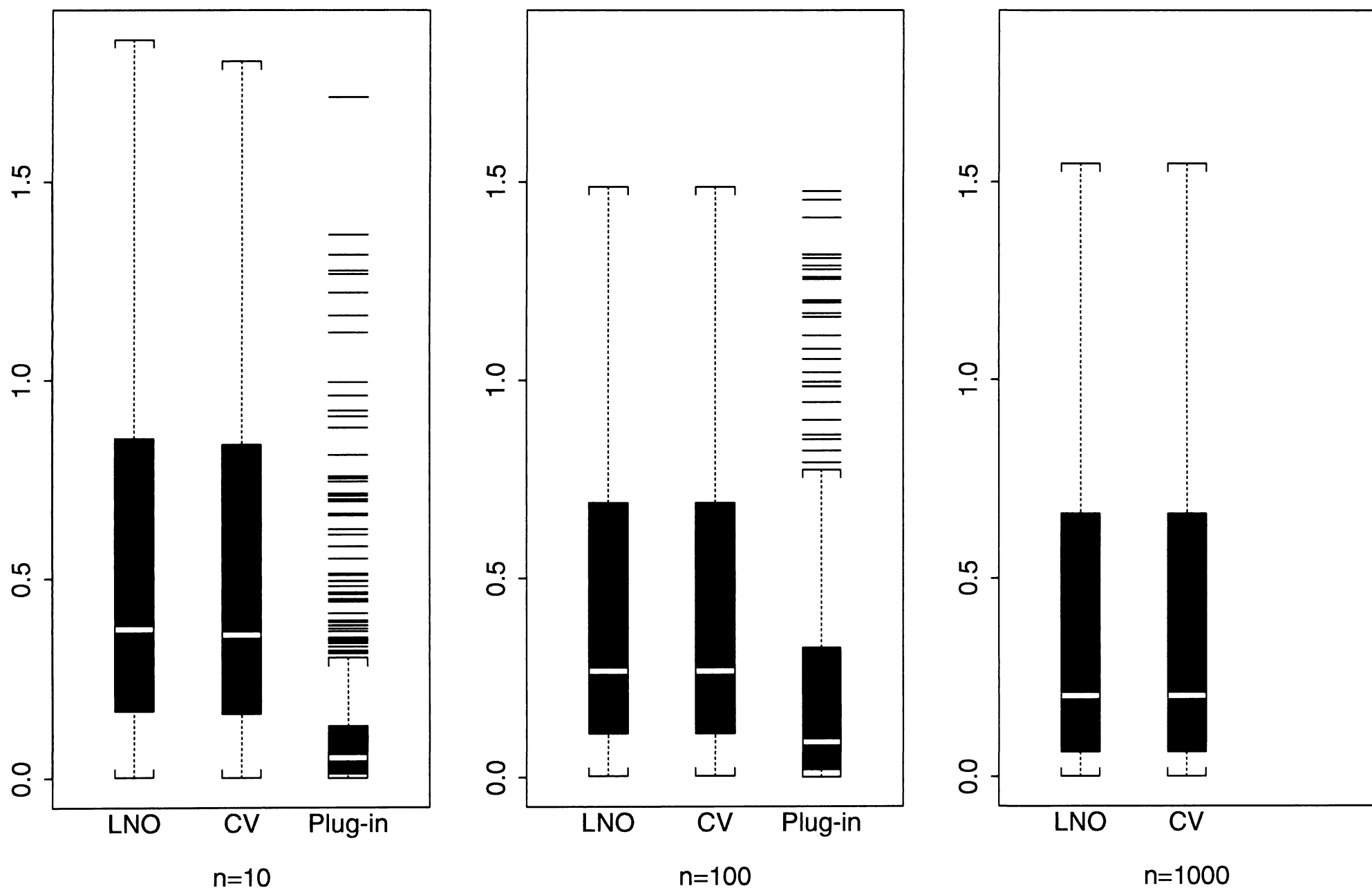


FIGURE 3: The ratio of adjusted ASEs (Equation 4.1) for each of LNO, CV, and Plug-in for 500 samples of size n from an exponential distribution. Notice that the plug-in estimator is the closest to optimal, and LNO and CV have similar behavior. Outliers greater than the largest whisker in each plot are not plotted. The weight function is $I[.1,3]$.